Localization Prediction of Proteins in Neutrophil Subcellular Compartments using Machine Learning

Yuhang Wang[1], Graziella Eliza Ronsein[1]

[1.] USP, Universidade de São Paulo, Av. Prof. Lineu Prestes, 748 - Butantã, São Paulo - SP;

Introduction - Predicting protein localizations in subcellular compartments can give clues regarding their functions. Neutrophils, which play a key role in the immune defense, have some granules types and their function. In this work, neutrophils were fractionated in Secretory Vesicles, Gelatinase, Azurophil and Specific compartments by ultracentrifugation. After that, fractions were analyzed by data dependent proteomics. Using the mass spectrometry results, we employed two machine learning training models to assist in predicting subcellular compartments. Material and method - After filtering for minimum of 3 valid values in at least one group of storage organelle, our data contained 368 proteins and 5 experimental replicates, each containing 4 compartments, totaling 20 columns with values of LFQ (label free quantification - obtained in Perseus). And that, we converted LFQ values into relative percentages for subsequent analyses. Firstly, we used Mclust in R package for unsupervised prediction, grouping proteins into clusters based on the experimental data. In this step, we calculated the correlation of replicates, and excluded proteins with a mean Pearson correlation below 0.7. After that, we applied t-SNE to reduce data\'s dimensionality to 2 and used Mclust function to group proteins into 12 clusters numbers. We assign the same compartments to proteins located in the same cluster as proteins with well-established subcellular compartments (based on extensive literature search). Secondly, we use Support Vector Machine (SVM) for prediction, a supervised method used to classify proteins based on previous literature knowledge. In this step, we analyze each replicate and find the best parameters (sigma and cost) for building the trainning model. After that, we apply svm classification to all proteins, obtaining their classification and score. With these scores, we determined threshold for eliminating proteins with low classification scores. Proteins were considered as belonging to a specific organelle if they had the same SVM classification passing the cut off threshold in at least 3 out of the 5 replicates. Result and discussion - Using Mclust function, we excluded proteins with a Pearson correlation value below 0.7 to eliminate non-reproducible experimental data. By comparing the 12 clusters with the literature, we assign 3 clusters for gelatinase, azurophil and specific (1 for each compartment) and 9 clusters for secretory vesicles (have richer proteome). As result, we obtained 32 proteins in Specific, 10 in Azurophil, 26 in Gelatinase granules, and 230 Secretory Vesicle. 70 proteins were not classified and remained with unknown location.For SVM, based on these scores after svm classification, proteins from each compartment with a score below the median were eliminated (classified as unknown). As result, for all replicates, we found 23 proteins in Specific granules, 13 in Azurophil, 16 in Gelatinase granules, 190 proteins belonging to Secretory Vesicle and 126 proteins with unknown location. Data divergence between the unsupervised and supervised approach may occur for proteins belonging to more than one compartment, due to experimental errors, and due to limitations of the computational methods. Conclusion - In this work, we were able to predict the localization of proteins in different neutrophil organellar compartments using two different machine learning methods. Comparison between methods increased the reliability of the predictions.

Agradecimentos: Agradeço principalmente à orientação da professora Graziella E. Ronsein para a confecção do presente trabalho. Ademais, agradeço igualmente a duas ex-alunas da professora: Gabrielly e Hellen, elas ofereceram suportes nos códigos propostos desse trabalho.